Automatic Translation to Controlled Medical Vocabularies

András Kornai[1] and Lisa Stone[2]

[1] Metacarta Inc, 875 Massachusetts Ave, Cambridge MA 02139
andras@kornai.com
[2] PPD Informatics/Belmont Research
84 Sherman St, Cambridge MA 02140

Abstract. In the medical domain, over the centuries several *controlled vocabularies* have emerged with the goal of mapping semantically equivalent terms such as `fever, pyrexia, hyperthermia,` and `febrile` on the same (numerical) value. Translating unstructured natural language texts or *verbatims* produced by healthcare professionals to categories defined by a controlled vocabulary is a hard problem, mostly solved by employing human coders trained both in medicine and in the details of the classification system. In this chapter we survey the automatic translation or *autocoding* systems currently in use.

0 Introduction

The current widespread use of controlled medical vocabularies emerged in response to the data exchange and standardization needs of modern medical research and care. In this chapter we survey the methods used in translating unstructured natural language texts or *verbatims* produced by healthcare professionals to categories defined by a controlled vocabulary.

To this day, the primary method of translation is to use human coders, trained both in medicine and in the details of the classification system. In the past few decades, automated translation systems or *autocoders* of varying efficiency and reliability have emerged. Our focus will be on system that are *hybrid* not only in the contemporary sense of including both rule-based and statistical inference techniques but also in the older sense of having both human and automatic components.

In Section 1 we describe the historical origins and current status of controlled vocabularies in the medical domain. In Section 2 we present the general architecture of hybrid autocoding systems, and discuss their main components.

1 Controlled medical vocabularies

Statistical methods pervade every aspect of medicine. The early detection of epidemics, the research methods determining the efficacy of drugs and other treatments, the diagnosis and treatment protocols used in actual patient care,

and even postmortem analysis are all based on careful statistical analysis of experimentally collected and naturally observed data points. But statistics can be meaningfully applied only if the source data is already properly classified. The need for controlled medical vocabularies to classify disease into general groups, and for detailed nomenclatures of signs, symptoms, diseases, and procedures has been recognized early on.

The City of London devised the *London Bills of Mortality* as an early warning system of the bubonic plague epidemics which periodically ravaged Europe. Beginning in 1603, these public postings provided London citizens with a detailed weekly mortality count, including cause of death and age of those fallen. After receiving a commission from King William III to study these records, John Graunt published his statistical analysis *Natural and Political Observations Made upon the Bills of Mortality* in 1662. This work is widely cited as the first known effort to classify human disease [23].

In the 18th century, as interest in natural history swept across Europe, physicians applied classification methods from botany and zoology to clinical medicine. They reasoned that diseases were entities like plants and animals and could likewise be arranged into taxomonic families, classes, species and genera on the basis of anatomical, clinical, and pathological criteria [28]. Francois B. de Sauvages, a professor of medicine at the University of Montpellier in France, published *Nosologia Methodica,* the first such systemic classification of disease in 1763. His system established 10 classes of disease, 44 orders, 315 genera, and approximately 2,400 separate entities. The scheme was enormously detailed but blemished by many inconsistencies and duplications [29]. Criticizing such cumbersome, all encompassing taxonomies, William Cullen of Edinburgh proposed a simpler arrangement in 1769. His *Synopsis Nosologiae Methodicae* was a didactic and practical index containing only 4 classes, 9 orders, and 151 genera [13].

The need for an internationally accepted classification system for statistical purposes and for public health control was recognized at the First International Statistical Conference in Brussels in 1853. The organization developed the *International List of Causes of Death.* This classification system, based on the work of the British statistician William Farr, had a two level hierarchy. The top level contained 5 general groups - epidemic diseases, constitutional diseases, local diseases according to their anatomical localization, diseases of the development, and diseases in direct consequence of a traumatism. At the lower level, 139 diseases were categorized. This classification was revised by the same organization in 1874, 1880, and 1886 [9].

In 1893, Dr. Jacques Bertillon, the chief statistician of Paris, resumed the work with the publication of *Nomenclatures de Maladies,* which became known as the Bertillon Classification of Causes of Death (WHO1) [54]. This system had a three level hierarchy. The highest level contained 44 groups, followed by a mid level containing 99 groups, and a lower level containing 161 disease entities. This system was adopted by many countries. The American

Public Health Association recommended its use for vital record registries. Upon Bertillon's death in 1922, the Health Organization of the League of Nations took responsibility for the system, publishing the fourth revision in 1929 [55], and the fifth revision in 1938 [56].

After WWII, the newly formed World Health Organization of the United Nations accepted responsibility for the decennial revision published in 1946. In 1948, at the sixth revision conference, the final document was published as *International Statistical Classification of Disease, Injuries, and Causes of Death.* This revision extended the scope of the classification to non-fatal diseases and added causes of morbidity to the causes of mortality. Most developed nations agreed to use this system for classifying and reporting public health statistics [9]. A 7th version containing minor revisions was published in 1955 [57].

1.1 ICD

In 1969, WHO published the 8th edition, and revised the name to *Manual of the International Statistical Classification of Disease, Injuries, and Causes of Death* (ICD). This revision contained many modifications designed to help statisticians index information from medical and hospital records [58]. In the United States, groups of physicians and health records administrators recognized that the 8th revision still did not meet their needs. As a result, the US government published the ICD-A, or the *Eighth Revision International Classification of Diseases, Adapted for Use in the United States* [11].

In 1977, WHO published the ninth revision of ICD, which was immediately adapted in the United States for clinical use as ICD9-CM. Most insurance companies still use this revision as a method for controlling billing and payment, although Version 10 was published in 1992 [59].

ICD9-CM is a hierarchical system containing 4 levels. In the 9th revision, there are 17 chapters at the highest level. A fixed range of 3 digit 2nd level codes is assigned to each of the 17 high level groups. Within each 2nd level range, a fixed range of codes is assigned to a midlevel grouping. The individual 3 digit code defines a concept at the mid (3rd) level, and two more digits are added to the code to classify the concept more exactly (4th or lowest level).

Since ICD assigns a 5-digit numeric code to each entry at the lowest level, the hierarchical organization serves as an important mnemonic aid to the human coder. It would be next to impossible to remember that 162.3 is the code for `malignant neoplasm of the upper lobe, bronchus or lung`, but let us trace the levels of this code through the ICD9-CM system. At the first level we find

1. Infectious and parasitic diseases (001-139)
2. Neoplasms (140-239)
3. Endocrine, nutritional and metabolic diseases, and immunity disorders (240-279)

4. Diseases of the blood and blood-forming organs (280-289)
5. Mental disorders (290-319)
6. Diseases of the nervous system and sense organs (320-389)
7. Diseases of the circulatory system (390-459)
8. Diseases of the respiratory system (460-519)
9. Diseases of the digestive system (520-579)
10. Diseases of the genitourinary system (580-629)
11. Complications of pregnancy, childbirth, and the puerperium (630-676)
12. Diseases of the skin and subcutaneous tissue (680-709)
13. Diseases of the musculoskeletal system and connective tissue (710-739)
14. Congenital anomalies (740-759)
15. Certain conditions originating in the perinatal period (760-779)
16. Symptoms, signs, and ill-defined conditions (780-799)
17. Injury and poisoning (800-999)

Obviously, the coder will select `neoplasms`, providing the range 140-239, for which we now inspect the 2nd level:

- malignant neoplasm of lip, oral cavity, and pharynx (140-149)
- malignant neoplasm of digestive organs and peritoneum (150-159)
- malignant neoplasm of respiratory and intrathoracic organs (160-165)
- malignant neoplasm of bone, connective tissue, skin, and breast (170-176)
- malignant neoplasm of genitourinary organs (179-189)
- malignant neoplasm of other and unspecified sites (190-199)
- benign neoplasms (210-229)
- carcinoma in situ (230-234)
- neoplasms of uncertain behavior (235-238)
- neoplasms of unspecified nature (239)

Thus from the 2nd level `malignant neoplasm of respiratory and intrathoracic organs` can be easily selected, providing the range 160-165. In this range, we inspect the 3rd level:

161 Malignant neoplasm of larynx
162 Malignant neoplasm of trachea, bronchus, and lung
163 Malignant neoplasm of pleura
164 Malignant neoplasm of thymus, heart, and mediastinum
165 Malignant neoplasm of other sites within the respiratory system and intrathoracic organs

Thus the coder can easily find `malignant neoplasm of trachea, bronchus, and lung`, providing the code 162. Going into the lowest (4th) level we find

162.0 Trachea
162.2 Main bronchus
162.3 Upper lobe, bronchus or lung

162.4 Middle lobe, bronchus or lung
162.5 Lower lobe, bronchus or lung
162.8 Other parts of bronchus or lung
162.9 Bronchus and lung, unspecified

and this provides the final (in this case, four-digit) code 162.3. In 2.3 we will see how this method of successive approximation remains applicable when we move from human to machine coding.

The 5 digit ICD9-CM codes are highly indexical in the sense that the numerical code can be analyzed to determine the position of a term in the hierarchy. While providing a terse mechanism for communicating information, indexical codes introduce many problems. If a disease entity is changed to a new position in the hierarchy, it must be assigned a different code. Also, it may be necessary to reuse obsolete codes to maintain the numeric hierarchy.

Reliance on a base ten numerical code results in limited flexibility and the need for "other", "not otherwise specified" (NOS), and "not elsewhere classified" (NEC) terms. Furthermore, the upper bound on the number of available codes limits the size of the vocabulary. Another problem is that the lack of multiple axes (independent classification dimensions) results in duplicate terms [12]. This was addressed in subsequent classification systems, in particular SNOMED, to which we turn now.

1.2 SNOMED

In the 1960s another organizational effort, arising from the need for greater specificity, and with the express goal of overcoming the limitations of ICD, was sponsored by the College of American Pathologists. This resulted in the publication of the *Systemized Nomenclature of Pathology* (SNOP) in 1965. SNOP was the first system to use the concept of multiple axes. The axes represent non-intersecting concepts which a medical condition could be classified under, such as anatomical site, environment, history, or etiology [10].

The system was primarily intended for the coding of surgical and autopsy diagnoses. The same organization published a successor, the *Systemized Nomenclature of Medicine* (SNOMED) in 1969, the same year that WHO published ICD9 Version 9. Over the last three decades SNOMED has been adapted by many organizations as a controlled terminology for the indexing of the entire medical record. The international SNOMED version 3.4 (October 1996) contains more than 150,000 terms distributed in 11 axes as follows:

1. Anatomy
2. Morphology
3. Function
4. Disease/Diagnosis
5. Procedures
6. Occupations

7. Etiology - Living organisms
8. Etiology - Chemicals
9. Etiology - Physical agents, forces, activities
10. Etiology - Social context
11. Etiology - Other links

Each axis has an arbitrary number of hierarchically organized terms beneath it. Like ICD, SNOMED assigns an indexical numeric code to each term. The first digit expresses a general entity, while each succeeding digit specifies a more detailed location in the hierarchy. Every concept is built up from a term from one or more axes: for example, `acute rheumatic fever` is identified with term code D3-17112, where D indicates the code is for the Disease Axis, 3 indicates a disorder of the cardiovascular system and 17113 indicates acute rheumatic fever. The code can be associated with codes along other axes. For example, along the topographic axis - T-15000, indicates cardiac disorder, and along the morphology axis - M-41000, indicates that the morphology is acute inflammation.

Since SNOMED contains some overlap of terms in different axes, it is possible to form two different versions for the same concept. For example, `acute appendicitis` has a single code, but there are also terms and codes for `acute`, `acute inflammation`, `appendix`, and `in`. Thus, the concept could be expressed either as `appendicitis, acute`, as `acute inflammation, in, appendix`, or as `acute, inflammation NOS, in, appendix`. This makes it difficult to compare similar concepts that have been indexed in different ways, or to search for a term that exists in different forms within a medical record.

While SNOMED permits single terms to be combined to create complex terms, rules for the combination of terms have not been developed. Consequently such compositions, such as `nevus of left esophagus` may not be medically valid [12]. Since SNOMED also uses indexical numeric codes, the same problems occur as with ICD when a term is relocated in a hierarchy.

The introduction of SNOP sparked intense debate between the proponents of statistical classification systems and those of more complex multiaxial medical nomenclatures. The former held that the primary uses of these systems was for international comparisons of disease and for cost reimbursement. For these purposes, only broad statistics on general groups and classifications of disease were needed. The proponents of multiaxial nomenclature maintained that single-axis systems were designed for general statistics and would become too complicated for a detailed nomenclature. Also, when several diagnostic codes are of interest, it is much harder to formulate complex boolean queries in a single axis than in a multiaxis code. Finally, the resolution offered by multiaxis codes could always be mechanically collapsed into a general-purpose single axis code, while the reverse operation is not feasible [9].

At the very time that these debates intensified, events occurred which abruptly shifted interest to statistical classification. The drug thaliomide caused thousands of severe birth defects in Europe and Canada, where it

had been recently approved, and in the United States, where it had been administered experimentally. As a result, sweeping regulations for pharmaceutical product development and adverse effect reporting were introduced [19]. The new regulations spawned a need for broader medical vocabularies for statistical classification of drug reactions and adverse events. Both WHO and the US Food and Drug Administration sponsored efforts to develop classifications systems specifically for this purpose.

1.3 COSTART and WHOART

The FDA published the *Dictionary of Adverse Reaction Terms* (DART) in 1967. DART was a hierarchical dictionary with the body-system category at the highest level. The FDA used DART for approximately 2 years, replacing it in 1969 with the *Coding Symbols for a Thesaurus of Adverse Reaction Terms* (COSTART). This new dictionary emphasized the specific adverse reactions over the system or organ affected, and allowed more than one search category to improve retrieval capabilities [45].

COSTART has a hierarchical arrangement of terms. At the top, there are 12 body-system categories. These are followed by subcategories and ultimately by coding symbols and preferred terms (a word or phrase of choice used to represent an adverse reaction). Instead of numbers, COSTART uses English language word symbols to record preferred terms. For example, the symbol ABDO PAIN codes `abdominal pain`. Coding symbols are also used for search categories (body-system classes) and their subcategories.

The same year, 1969, saw the publication of the *WHO Adverse Reaction Terminology* (WHOART) by the World Health Organization. WHOART is also hierarchical, with 30 system-organ classes followed in the hierarchy by high-level terms and preferred terms. WHOART assigns a numeric code to all system-organ classes, high-level terms, and preferred terms. Since WHOART was published in four languages (English, French, German, and Spanish), the use of non-language specific codes was an important feature which distinguished it from COSTART [45].

For the next quarter century, adverse event coding was dominated by WHOART (required by the European Union) and COSTART (mandated by the FDA). The FDA translated all adverse reactions reported to the agency into COSTART terminology [51]. To report to the World Health Organization, the FDA converted the COSTART codes into WHOART codes, using a fixed translation table [45]. In combination with an adverse reaction terminology, most organizations processing regulatory data also used a morbidity terminology - primarily ICD-9 in Europe and ICD9-CM in the United States. The Japanese developed their own versions of these international terminologies, *Japanese Adverse Reaction Terminology* (JART) and MEDIS [39,47].

Over the years, organizations expressed dissatisfaction with these established terminologies. Deficiencies included a lack of specificity at lowest level of terms, limited data retrieval options (e.g., too few levels in the hierarchy,

or capacity to retrieve data via one axis only) and ineffective handling of syndromes. Reassignment of codes in new versions forced existing data to be reevaluated. Use of one terminology for adverse reactions, and other for morbidity complicated data retrieval and analysis [39].

To address these deficiencies, some organizations modified the existing terminologies for company or regional use, or developed entirely new systems. Some of these, in particular the *Hoechst Adverse Reaction Terminology System* (HARTS), spread well beyond the parent company (Hoechst, now Aventis). This created additional problems. The use of different terminologies in separate geographic regions impaired international communication and necessitated the cumbersome conversion of data from one terminology to another. These problems most acutely hindered multinational pharmaceutical companies whose subsidiaries used multiple terminologies to fulfill the varying data submission requirements of various national regulatory agencies. As pharmaceutical companies merged and became global, it became increasingly difficult to manage the information required for product development and regulation.

1.4 MedDRA

In 1994, the International Conference on Harmonization (ICH) introduced multi-disciplinary regulatory communication initiatives to compliment their ongoing safety, quality, and efficacy harmonization efforts. This included the M1 initiative which sought to standardize international medical terminology in all phases of the regulatory process. The initiative resulted in the development of the *Medical Dictionary for Regulatory Activities* (MedDRA), based on the UK Medicines Control Agency's (MCA) medical terminology. Version 2, the first implementable version, was published in 1997. Since then, new releases have appeared regularly and MedDRA has rapidly been adopted an internationally accepted medical terminology for regulatory purposes. The FDA already recommends (though does not yet require) MedDRA, and the European Union will require MedDRA coding for submissions and safety reporting after January 2003 [39]. MedDRA is a multiple axial system that uses non-indexed codes, and includes a five level hierarchy of of terms.

Table 1. MedDRA terms in the 4.0 version

Level	Name	# terms
SOC	System Organ class	26
HLGT	High Level Group Term	334
HLT	High Level Term	670
PT	Preferred Term	15000
LLT	Low Level Term	55000

The SOC level is roughly analogous to the highest levels of COSTART and WHOART. The HLGT and HLT levels group related terms beneath them by anatomy, pathology, etiology or function. The PT level contains distinct descriptors medical concepts including symptoms, signs, diseases, procedures and social characteristics. The LLT level contains terms which are synonyms or lexical variants of the PT to which they are linked. Since LLTs accommodate culturally unique terms, each LLT may not have a translation in all languages. An LLT can be linked to only one PT.

Let us trace the term `malignant neoplasm of the upper lobe` through the five levels of the MedDRA system. At the SOC level the term may be classified either on the the etiological axis provided by `Neoplasms benign and malignant`, or the or the the anatomical axis provided by `Respiratory, thoracic and mediastinal disorders`. Selecting the former leads to the HLGT `Respiratory and mediastinal neoplasms malignant and unspecified`, the HLT `Respiratory tract and pleural neoplasms malignant cell type unspecified NEC`, the PT `Lung cancer stage unspecified` and finally the LLT `malignant neoplasm of upper lobe`.

If we select the anatomical axis, we begin with the SOC `Respiratory, thoracic and mediastinal disorders`, the HLGT `Respiratory tract neoplasms`, the HLT `Lower respiratory tract neoplasms` and the PT `Lung cancer stage unspecified`. Here we see that the PT was included under two HLT terms. While a PT can have more than one associated HLT (and HLGT, and SOC) there is at most one path down through the hierarchy (SOC-HLGT-HLT-PT) to any PT, so that counting errors are avoided when querying on the SOC, HLGT, and HLT levels. MedDRA also assigns a primary SOC (and therefore a primary path PT-HLT-HLGT-SOC) for consistency in standard reporting, based on most common usage.

1.5 Other terminologies and systems

While MedDRA looms large in the regulatory landscape, and can be said to subsume several of the earlier systems, including COSTART, WHOART, and HARTS, our survey would not be complete without describing at least the most widely used systems that currently fall outside its scope.

DSM Perhaps the largest medical field with its own well-developed terminology is psychiatry. The American Psychiatric association first published its *Diagnostic and Statistical Manual of Mental Disorders* (DSM) in 1952. The most recent edition, DSM-IV was published in 1994, with minor modifications in 2000. This system, which has been adopted by many countries, has five axes:

1. Clinical Disorders
2. Personality Disorders and Mental Retardation
3. General Medical Conditions

4. Psychosocial and Environmental Problems
5. Global Assessment of Functioning Scale

Beneath the axes, diagnostic labels and their corresponding codes are grouped. Codes contain 3-5 digits. The leading 3-4 digits specify a specific entity, with the 5th digit providing additional specificity such as subtype or severity. To provide compatibility for reimbursement and other interested third parties, each DSM-IV code is associated with a ICD code equivalent [15].

ICPC In 1972, the World Conference of Family Doctors initiated a project to develop a comprehensive system to classify the primary health care and patient-doctor interactions. A succession of efforts led to the publication of the International Classification of Primary Care (ICPC) in 1978. The second edition was published in 1998, primarily to coordinate the codes with the the 10th version of ICD [61]. Several European countries including the Netherlands, Denmark, and Norway have produced national ICPC implementations [26].

The ICPC system groups codes under a matrix of 17 chapters and 7 components. The chapters are:

A General and Unspecified
B Blood, Blood Forming Organs and Immune Mechanism
D Digestive
F Eye
H Ear
K Circulatory
L Musculoskeletal
N Neurological
P Psychological
R Respiratory
S Skin
T Endocrine, metabolic, and nutritional
U Urinary
W Pregnancy, child-bearing, family planning
X Female genital and breast
Y Male genital
Z Social problems

These are roughly analogous to the highest level in WHOART, COSTART, MedDRA, etc. The components, which offer a different organizational principle, are as follows:

1. Symptoms and Complaints
2. Diagnostic Screening and Preventative Measures
3. Medication and Treatment procedures

4. Test results
5. Administration
6. Referrals and other reasons for encounter
7. Diagnoses and Disease

Components 2-6 are referred to as *process* components. Codes are grouped under a Chapter and a Component. Codes consist of three characters and have a title of limited length. Each code is mapped to one or more corresponding ICD-10 codes. A patient-physician interaction is assigned a series of codes, for example `A03 - Fever, A30 - Full examination, A31 - Temperature Measurement, A63 - Plan for follow-up visit.`

Read Originally created by Dr James Read to generate computer based morbidity records, the Read Classification Codes (RCC) were adopted by British National Health Service in 1994, and were expanded to to cover all fields of primary health care with the publication of Version 3.

This version includes a 5 level hierarchy and allows classification of terms along multiple axes. The non-indexical 5 character codes are cross-referenced to ICD10 codes. In Version 3.1, the current version, a set of qualifier terms such as anatomical site was added. Qualifier terms can be combined with existing terms to form composites which exist outside of the hierarchy. Terms and qualifiers are grouped into templates that describe the range of medically valid combinations. Many medical organizations, primarily in the United Kingdom, use the RCC for classification of medical records in clinical information systems [41].

UMLS The proliferation of controlled vocabularies has long worried information retrieval specialists, who would like to cross-reference information coded according to different schemes. The National Library of Medicine is developing the *Unified Medical Language System* (UMLS) with the specific goal of addressing these issues [52]. However, the UMLS Metathesaurus is nowhere near completion.

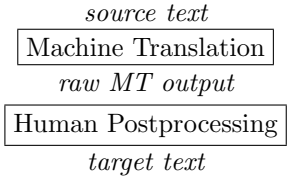
Given the diverse constituencies of the various schemes, it is highly unlikely that a comprehensive uniform scheme will be adopted within a decade or two. This simple fact has major impact on the design of hybrid translation systems: since any such system needs to "have legs" (require no or little external effort to move it from one scheme to another), it makes little engineering sense for them to contain detail knowledge about the internal organization principles of any particular coding scheme, and even less sense to tie the behavior of the system to the semantics of such schemes. In other words, the fact that we are considering a narrow domain, medicine, does in no way imply that we can, or even should try to, utilize deeply domain-specific knowledge. To the contrary, in order to keep the system re-targetable for different controlled vocabularies, we should rely on as generic machine translation and machine learning techniques as we can. For this reason, in the following we

put the emphasis on generally applicable techniques as opposed to systems that are tied very closely to any particular target vocabulary.

2 Hybrid Autocoding Systems

The ability to master natural language communication played a central role in the original Turing test [50], and for well over a decade Machine Translation (MT) was believed to be right around the corner. In spite of some very clear early warnings like [2], it was not until the appearance of the ALPAC report [21], that the actual difficulty of the problem became widely appreciated. But once fully automated high quality MT was recognized to be the extremely hard problem it is, attention turned to restricted/reformulated versions. As a first step, "high quality" was replaced by "with some effort understandable" and "fully automated" was replaced by "suitable for human post-processing". Thus the data flow followed the scheme outlined in Fig. 1:

Figure 1. Machine-assisted translation (coding) workflow



Another important change was the move from unrestricted to domain-specific text. In fact, for many domains where the syntax is highly conventional and the content varies little, e.g. in weather reports [7], the labor-intensive post-processing step could be kept to a minimum. In general, interactive translator's workbench systems, where the results of the postprocessing step are continually fed back to the MT component, were found more effective than single-pass postprocessing.

It was soon realized that the success of such systems derives largely from the conventionalized, repetitive aspect of the task, and that conventionalization is best accomplished by controlling the source language. This is a very active research area to this day, but one that we can't do full justice to within the bounds of this chapter, and we refer the reader to the survey [60] and the proceedings of the biannual CLAW conferences published by the Language Technologies Institute at Carnegie Mellon University.

From the perspective of MT, the main distinguishing feature of the medical domain is the size of the vocabulary. The METEO system [7] of translating weather reports could be successful precisely because it knows less than 10^3 words, and it is possible to handcraft lexical entries reflecting the usage possibilities of each and every one of these. But even the tightly controlled vocabularies surveyed in the previous Section contain $10^4 - 10^6$ words and

phrases, and our ability to handcraft meaningful lexical entries for each of them is limited.

Over the years, many researchers have noted that a great deal of the (controlled) vocabulary is *compositional* in the sense that the meaning of the whole expression can be recursively inferred from the meaning of its parts. To take advantage of this fact, one of course needs to define the meaning of the expressions in some knowledge representation scheme. A significant effort in this direction is GALEN, now OpenGALEN, which builds a whole conceptual model of the medical domain, with knowledge representation primitives such as ACTS_ON, IS_PART_OF, etc.

Since it is clearly the long-term goal of medical informatics to use such representations as the basis of inference, we include a discussion of the problems and main techniques here, but we caution the reader that the representation of medical knowledge is still in its infancy, with narrow, non-scalable prototypes dominating the field, while the use of autocoders is part of the mainstream production environment for pharmaceutical research and regulatory activities.

We distinguish three interrelated problems in the translation process, and discuss them in turn:

1. *Segmentation,* the problem of finding (the beginning and endpoints of) the words and phrases that are subject to translation
2. *Analysis,* the problem of identifying and recursively substituting smaller patterns that compositionally make up the larger patterns and
3. *Substitution,* the problem of providing the appropriate translation.

While separating the tasks at the conceptual level is clearly necessary, it should be emphasized that a working system need not contain separate modules for each, nor is it necessary to impose an execution order in which solving one of these problems precedes solving the other. In fact, lack of a rigid pipelined architecture, and the ability to delay decisions, is a key strength of the most important hybrid methods.

2.1 Segmentation

Finding the relevant words and phrases is a problem long familiar from machine translation. It is particularly acute in languages such as Chinese or Japanese where no whitespace separates the symbols, and in languages such as German where long compound words are common. Current segmentation and *entity extraction* techniques have their historical roots in string matching techniques – for a survey, see [46]. The key idea, built into Unix in the form of `grep` [49] is to describe string patterns in terms of regular expressions, and to speed up regular expression searches by means of compiling them into finite automata.

To check for an entity, the current generation of autocoders perform some combination of the following abstraction steps [22,18,40]

1. special character normalization – replacing # by `number`, @ by `at`, etc
2. spell correction – replacing `pian` by `pain`
3. acronym expansion – replacing `ECG` by `electrocardiogram`
4. undoing abbreviations – replacing `liv` by `liver`
5. capitalization – lowercase chars replaced by their uppercase counterpart
6. morphological analysis (stemming) – `fingers` replaced by `finger`
7. synonym-based data enrichment – replacing `ache` by `pain`
8. word deletion – omitting words low in information content

Notice that all these steps correspond to *generalized sequential mappings* that map input strings on output strings in an essentially memoryless fashion. The idea of "memoryless" computation may be somewhat counterintuitive for steps such as acronym expansion or data enrichment, but in fact dictionary entries can be thought of as containing two strings, an input and an output, and looking up a string in the dictionary is an operation that requires no writable memory. Such operations can be performed by a particularly simple kind of computing machinery, *finite state transducers* or FSTs. The techniques for compiling word lists into FSTs are well understood, and have been widely used in tokenization tasks since the introduction of `lex` [35].

In the eighties, the whole technology of string matching received a new impetus from the work of Koskenniemi [33], who showed how simple FSTs operating in parallel can be used to express complex relations between strings. Since both parallel and sequential combinations of FSTs can be compiled into single (albeit often much larger) FSTs, the road was open to express regularities by the kind of rules used by lexicographers and grammarians, compile the rules into FSTs, and combine the FSTs into a single, highly efficient program that will check for all the regularities at the same time. Crucially, the computational efficiency of the resulting FST is the same whether it is used in synthesis (generation from input to output) or analysis (recognition of input from output) mode. For more detailed discussion of these techniques see [42,31]. In autocoding, and perhaps in all of machine translation, probabilistic FSTs are the single most important hybrid technique, since they enable the compilation of expert-written and machine-generated rules in a unified framework.

In spite of their mathematical similarity, not all the above steps are equally practical. Since special character normalization is the key to successfully matching certain entities, and the cost of implementing this step is negligible, it is best included in any system, but on the whole will affect only a small fraction, well below .1%, of the verbatims to be translated. Spell correction, on the other hand, affects a larger fraction, often as much as 1-2%, but general-purpose spellcheckers perform poorly in the medical domain, and implementing a domain-specific spellchecker is a major undertaking. Morphological analysis (stemming) yields a small but consistent improvement of 4-5% in autocoding, which is not nearly as good as that found in [34] for the

most comparable situation in Information Retrieval (IR), short queries and short documents. The use of data enrichment is highly controversial, because the errors it can introduce, being semantically valid, are hard to spot. On the whole, we feel that this step should be performed indirectly, on the output knowledge representation, rather than directly, on the input dictionary words. Omitting *stopwords* that have low information content is very effective, but care must be taken in selecting the stopword list. For a summary of the effectiveness of these steps for IR see [14].

There is one widely used technique, word order normalization, which does not easily lend itself to reformulation in terms of FSTs. The idea, closely related to the "bag of words" model long familiar from information retrieval, is to simply list the words in some fixed (typically, alphabetical) order. By this method, many equivalent expressions such as `enlarged left kidney`, `left kidney, enlarged,` or `kidney, left enlarged,` can be mapped on a single normalized order, in this case `enlarged, kidney, left`. As a practical matter, word order normalization creates almost as many problems as it solves, since mechanical application of the procedure will map medically distinct conditions such as `pain without swelling` and `swelling without pain` on the same form `pain, swelling, without`.

Early entity extraction systems were entirely rule-based, with rules such as (paraphrase) "the word or phrase following `normal` or `abnormal` refers to a test result or body function". Such rules suffered from all the problems of classic expert systems: they were

- *expensive* – could only be generated by experts well versed both in the subject domain and in the intricacies of rule writing
- *brittle* – minor changes in the data often require radical revisions of the rule system
- *fragmentary* – relevant data may not be found even in large samples, and
- *monolithic* – changes cannot be effected in a modular fashion, every rule can interact with every other rule, to produce unexpected changes in overall behavior

With the introduction of statistical machine translation techniques [6], entity extraction became less of an art and more of a practicable technique. While such techniques address the cost, brittleness, and monolithicity issues, the problem of fragmentary data remains, and it has been the general experience that humans are far more capable of extracting the relevant patterns from fragmentary data than the currently known statistical techniques.

Therefore, the best systems still have a strong handcrafted/human supervised component, and the expertise of human coders still very much needs to be tapped into both at the stage of preparing the dictionaries and synonym tables and at the stage of assuring the quality of the machine output. How this "hybridization" itself can be performed automatically or semi-automatically will be discussed in 2.3.

2.2 Analysis

To a large extent, the task of finding the entities to be translated is intervowen with the parsing of the text: very often, it is not the first word or stem of the entity itself that provides the necessary clue, but rather the preceding word, and the same is true about detecting the end of an entity. That said, obtaining a full parse is seldom a necessary prerequisite to locating the phrases to be translated, and indeed the complexity of the full grammatical task is formidable, even in the medical domain, which itself is rather restricted compared to general-purpose English. Here we list some of the most salient properties that set adverse event coding apart from the bulk of the work on machine translation.

Table 2. Medical vs. general-purpose translation

characteristic	adverse event coding	general MT
length	short (on average < 6 words)	long (multi-paragraph)
syntax	nonstandard `acne facial recurrent`	standard `recurrent facial acne`
language	mixed over 10% German, French, etc.	monolingual English
style	telegraphic `allergy arms and legs`	normal `skin allergy on the arms and legs`
vocabulary	specialized 10^5 words and phrases added to	general-purpose 10^5 words in gen. vocabulary

Most of these differences extend well beyond adverse event coding: similar issues need to be faced in coding to the disease, symptom, and procedure classifications surveyed in Section 1, and in some cases, such as coding to drug names, specialized vocabulary may grow into the millions. To some extent, the above characteristics are correlated: the longer the text grows, e.g. in discharge reports, the more it sheds from its nonstandard properties and the more it becomes like general-purpose English. This would be an advantage if we had at our disposal high-quality parsers for general-purpose English just the way we have high-quality spellcheckers. Unfortunately, such parsers are beyond the state of the art, and we only see the downside of moving closer to ordinary English: more mistakes and ambiguities, a wider range of syntactic constructions, and an overall loosening of the implicit language control that is imposed by the rigid constraints of the genre.

Because of these factors, the analysis of medical text tends to follow, rather than lead, the state of the art in analyzing general-purpose texts. There the most important technique to have emerged in the past decade is *shallow parsing* or *chunking* [1], which abandons the goal of providing a full analysis, and views the extraction of grammatical categories such as Noun

Phrase or Verb Phrase as an extension of the entity extraction task, relying on the same general finite-state techniques we discussed in 2.1 above. An important consequence of this trend is that the output of parsing (syntactic analysis) is no longer sufficiently detailed to provide input for knowledge representation (semantic analysis), and the words and phrases extracted from the text are substituted in more loosely defined semantic patterns [37]. This becomes particularly noticeable when we encounter medically meaningless expressions such as `removal of a renal cyst from the thyroid` [43]: while in principle a sophisticated knowledge representation scheme as found in OpenGALEN may be able to handle these correctly, exercising this capability over a wide range of constructions presupposes a level of parsing accuracy that is currently not available.

A relatively easy step is sentence- and clause-level parsing, i.e. to identify chunks separated by hyphens, parentheses, and other punctuation. For autocoding, as in the WebCoder system [32], this has the advantage that the chunks can be ranked according to their hypothesized hierarchical importance. Chunks shorter than four words that constitute the whole verbatim are considered *definitive* in the sense that they contain all the information necessary for substitution and receive rank 0, partial chunks have rank 1 and higher. As Table 3 shows, the number of less valuable (higher ranked) chunks emerging from a shallow parser falls of rapidly with rank. The quality of the information contained in lower-ranking chunks is also rapidly decreasing [32].

Table 3. Distribution of chunk ranks

	WHOART	*WHOART*	*COSTART*	*COSTART*
	tokens	types	tokens	types
trainset	64765	64515	50301	8038
testset	7195	7195	5589	1718
rank 0	62636	62366	50264	7946
rank 1	10759	8630	1796	556
rank 2	649	524	92	38
rank 3	22	22	2	2
rank 4	2	2	0	0

2.3 Substitution

Originally, the computational effort of autocoders was dominated by database lookup. These were little more than *translation memory* (TM) systems (for an overview, see [53]) that simply stored previously encountered entities with their proper encoding, as assigned by the human coder. The effectiveness of TM is almost entirely a function of the size of the translation memory or *synonym table* as it is called in autocoders. Since the tables are carefully constructed and debugged by human coders familiar with the coding schemes surveyed in Section 1 [17], their error rates are very low, but they miss a

great deal. Today, in translating to a controlled medical vocabulary, the key problem lies not so much with providing the actual translation (which can be accomplished by standard hashing/database lookup methods), but with filling the synonym table with minimum human effort. Our interest is not so much with memorization as with Machine Learning (ML) techniques that can generalize better to data not seen in training, even at the cost of increased error.

Here we define two figures of merit standardly used in computational linguistics: by the *precision* of the system we mean the percentage of correct output relative to the total output, and by *recall* we mean the percentage of correct output relative to total input. How input and output is defined depends on the nature of the task, such as entity extraction, information retrieval, classification, or machine translation, but the intent is always the same: the precision measure is defined to be sensitive to false positives, while the recall measure is defined to be sensitive to false negatives. In practical applications, it is almost always possible to trade off precision for recall and conversely. TM systems are at one extreme of this tradeoff: precision is very high, but recall can be as low as 15%, and 40% is considered very good [5].

Rather than thinking of substitution as a mechanical database lookup step, it will be expedient to recast the whole problem as an instance of the general IR/text classification problem, or more specifically, as an instance of the message routing (MR) problem [24]. In supervised MR we are faced with the following problem: given some categories $C = \{C_1, C_2, \ldots, C_k\}$ and some "truthed" documents $F = \{F_1, F_2, \ldots, F_n\}$ with their true categories $t(F_i) \subset C$, for any new document F_{n+1}, find the most likely value of $t(F_{n+1})$, i.e. the category or categories associated to F_{n+1}. In a typical MR system files are first reduced to multisets or *bags* of words or word stems in the preprocessing stage, so that the task becomes classifying these bags of words, or stems, rather than the original files. Some systems, such as [32], go beyond words and employ word pairs and in general word n-grams, since these are more informative than word unigrams alone.

Given an arbitrary fixed ordering of words/stems, e.g. as created by a perfect hash function, bags are in one to one correspondence with vectors of a large dimensionality d (the number of words, usually several thousand to a few hundred thousand) having only nonnegative integer coefficients called in these applications the *counts*. Looked at this way, autocoding (and in general, supervised MR) is a classification problem in Euclidean space. Before turning to a discussion of the main machine learning approaches, however, we list the main aspects of the MR problem that set it apart from other important tasks such as Optical Character Recognition.

(1) The number of classes is large, often $10^4 - 10^5$
(2) The number of potential features (word n-grams) is astronomical
(3) Human judgments are available
(4) The data is repetitious

All machine learning algorithms solving supervised problems have two phases of operation: *training* and *classification*. In training, a mathematical model of the classes is created based on pairs consisting of verbatims and their translations. In classification, the model is applied to new data, and hypothesized translations (and confidences) are output. With some simplification, we can distinguish two broad types of training methods: those that create a separate model for each class, and those where knowledge about the individual classes is distributed over the whole system.

All handcrafted classifiers, based on (combinations of) keywords and rules, fall in the first type, where we also find automatically generated models such as Naive Bayes [38] and linear machines [25,36]. The second type includes many variants of Artificial Neural Nets (ANNs) [44] and k-nearest neighbor classifiers (kNN). The string normalization techniques discussed in 2.1 are for the most part very conservative, assuring only that near-identical strings receive the same translation. We can therefore think of Translation Memory systems as nearest neighbor classifiers, but the abstract string-similarity neighborhoods around the input strings listed in the synonym table are very small.

Within the bounds of this survey we can't do full justice to all these techniques: for an overview of the main machine learning methods see [16], and for the state of the art in text classification see the annual SIGIR and TREC proceedings. But we emphasize here that much of what we know about Message Routing comes from relatively small collections of data, such as the original Reuters corpus, which has less than 30k documents and only 90 categories with both training and test examples. Almost all this knowledge needs to be critically reevaluated for autocoding. Because of (1), techniques that are not linear or at least near-linear in the number of classes are essentially useless even when they are provably superior on small data sets: this includes much of standard multivariate statistics, such as factor analysis [4] and regression methods [3].

One well-understood technique for dealing with this problem is divide and conquer: we can leverage the hierarchical structure of the classes by first classifying to the highest level, and building independent lower-level classifiers for each node. In [32] we built a bodysystem-level master, with bodysystem-specific slaves, at the cost of increasing reject rates by 1.3%. and error rates by 0.1%. It has been argued that in some cases no loss (or even accuracy gains) could be seen [30], but these results may depend on experiments being performed in a more noisy environment than provided by medical texts. Because of (3), we have much better training data than is customary in information retrieval: essentially, all previously coded material, and all preexisting synonym tables, are at our disposal. Also, these are very high quality both in terms of extremely low error rates (often coded twice by different coders and reconciled by a third coder if needed) and high consistency (mature coding

practices). Laveraging this training material properly is the key to the success of hybrid systems [27]

Because of (2), feature selection can matter a great deal. A widely used technique is Singular Value Decomposition (SVD) [8,20], which provides good dimension reduction, but the nonlinear computational cost is almost impossible to absorb for very large problems [48]. Since the same drug is likely to have the same adverse effect on different people, we also benefit from (4) – with the proper statistical techniques both (3) and (4) can be highly leveraged to mitigate the effects of (1) and (2).

3 Conclusions

Automatic translation to controlled medical vocabularies is not a solved problem. Because of the characteristics of the medical domain, many techniques that work well for less demanding data sets are not practical for autocoding, and properly leveraging the knowledge of human coders remains the key to developing and deploying successful systems.

Acknowledgment

The work on WebCoder was supported under SBIR grant # 2 R44 CA 65250 from the National Cancer Institute, National Institutes of Health. The authors would like to thank Michael Johnston, Jeremy Pool, and J. Michael Richards for their help at various stages of the project.

References

1. Steven Abney. 1991. Parsing by chunks. In Robert Berwick, Steven Abney, and Carol Tenny, editors, *Principle-based parsing*. Kluwer Academic Publishers.
2. Bar-Hillel, Y 1960 A demonstration of the nonfeasibility of fully automatic high quality translation. In FL Alt (ed): *The present status of automatic translation of languages* Academic Press, New York, 158–163
3. Peter Biebricher, Norbert Fuhr, Gerhard Lustig, Michael Schwanter, and Gerhard Knorz 1988 The automatic indexing system AIR/PHYS – from research to application 11th Int Conf on R&D in IR 333-342
4. Harold Borko and Myrna Bernick 1963 Automatic document classification. JACM 10 151–161
5. Sonja Brajovic 2002 Personal Communication
6. PF Brown, SA Della Pietra, VJ Della Pietra, and RL Mercer 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* **19** 263–311.
7. M. Chevalier, J. Dansereau, and Poulin, G. 1978. TAUM-METEO: Description du Système. Technical report, Groupe de recherches pour la traduction automatique Université de Montréal

8. Christopher G. Chute and Yiming Yang. 1995. An overview of statistical methods for the classification and retrieval of patient events. *Methods of Information in Medicine*, 34:104–109.
9. Roger A. Cote, and Stanley Robboy. 1980. Progress in Medical Information Management, Systemized Nomenclature in Medicine (SNOMED). *Journal of the American Medical Association*, 243:756–762.
10. Roger A. Cote. 1978. The SNOP-SNOMED Concept; Evolution Towards a Common Medical Nomenclature and Classification. *Proceedings of the Seventh International Congress on Medical Records*
11. Cimino JJ., G Hripcsak, S Johnson and P Clayton 1989 Designing an introspective, controlled medical vocabulary. In: Kingsland L, ed. *Proceedings of the 13th annual symposium on computer Applications in Medical Care*. IEEE Computer Society Press, Washington, DC
12. Cimino JJ 1996 Review paper: coding systems in health care. *Methods Inf Med*, 35(4-5):273–284.
13. William Cullen. 1792 *Synopsis and Nosology* N. Palten Publishers Hartford, CT
14. Guy Divita, Allen C. Browne, and Thomas C. Rindflesch. 1998. Evaluating lexical variant generation to improve information retrieval. In *Proc. American Medical Informatics Association 1998 Annual Symposium*, Orlando, Florida.
15. American Psychiatric Society 2000. *Diagnostic and Statistical Manual of Mental Disorders DSM-IV, 4th Edition Text Revision* American Psychiatric Society, Washington, DC
16. Richard O. Duda, Peter E. Hart and David G. Stork 2000 Pattern Classification Wiley, New York, NY
17. Thérèse Dupin-Spriet and Alain Spriet. 1994. Coding errors: classification, detection, and prevention. *Drug Information Journal*, 28:787–790.
18. Christian Fizames. 1997. How to improve the medical quality of the coding reports based on WHOART and COSTART use. *Drug Information Journal*, 31:85–92.
19. Food, Drug, and Cosmetic Act 1962. Kefauver-Harris Amendment, 21 U.S.C. Section 355
20. Stephen I. Gallant. 1995. Exemplar-based medical text classification. *Belmont Research SBIR Proposal 1 R43 CA 65250-01.*
21. Paul Garvin 1966. Language and Machines. Computers in Translation and Linguistics. (ALPAC report, 1966). National Academy of Sciences.
22. Terry L. Gillum, Robert H. George, and Jack E. Leitmeyer. 1995. An autoencoder for clinical and regulatory data processing. *Drug Information Journal*, 29:107–113.
23. John Graunt 1662. *Natural and Political Observations Made Upon The Bills of Mortality London.*
24. Donna K. Harman, editor. 1994. *The Second Text REtrieval Conference (TREC-2).* National Institute of Standards and Technology, Gaithersburg, Maryland.
25. Wilbur H. Highleyman. 1962. Linear decision functions with application to pattern recognition. *Proceedings of the IRE*, 50:1501–1514.
26. Marc Jamoulle. 2001 ICPC use in the European Community *16th WONCA World Congress of Family Doctors* Durban, South Africa Available at http://www.ulb.ac.be/esp/wicc/icpc_2001.html, accessed March 12, 2002

27. Michael. C. Joseph, Kathy Schoeffler, Peggy A. Doi, Helen Yefko, Cindy Engle, and Erika F. Nissman. 1991. An automated COSTART coding scheme. *Drug Information Journal*, 25:97–108.
28. Lester S. King. 1966 Boissier de Sauvages and 18th century nosology. *Bulletin of Historical Medicine*, 40: 43–51
29. Lester S. King. 1958 In *Medical World ofthe Eighteenth Century* , Chapter 7: 193-226 University of Chicago Press, Chicago
30. Daphne Koller and Mehran Sahami 1997 Hierarchically classifying documents using very few words In *Proc. 14th Int Conf on Machine Learning* 170–178
31. András Kornai 1999. Extended finite state models of language Cambridge University Press, Cambridge, England
32. András Kornai and J. Michael Richards. 200 In A. Abraham, M. Koeppen (Eds.) *Hybrid Information Systems*: 527–538 Physica Verlag, Heidelberg
33. K. Koskenniemi 1983. Two-level morphology: a general computational model for word-form recognition and production. Ph.D. thesis, University of Helsinki.
34. Robert Krovetz. 1993. Viewing morphology as an inference process. In *Proceedings of SIGIR93*, pages 191–202.
35. Michael E. Lesk 1975 Lex — a lexical analyzer generator. Technical Report 39, AT&T Bell Laboratories, Murray Hill, NJ,
36. David D. Lewis, Robert E. Schapire, James P. Callan, and Ron Papka. 1996. Training algorithms for linear text classifiers. In *Proceedings of SIGIR96*, pages 298–306.
37. David M. Magerman 1995. Statistical Decision-Tree Models for Parsing. In *Proceedings of ACL95*, pages 276–283
38. ME Maron 1961 Automatic Indexing: an experimental inquiry. JACM 8 404–417
39. MSSO-DI-6003-4.1.0 2001 *MedDRA Introductory Guide, version 4.1*
40. Keya Pitts, Kathleen Jelliffe, and and Lionel Benson 1999. Advances in Volume Encoding Clinical Data. *Drug Information Journal* 33/4: 1079–1092.
41. NHS Centre for Coding and Classification. 1993 Read codes and the terms projects: a brief guide. London: Department of Health
42. E. Roche and Y. Schabes 1997. Finite-State Devices for Natural Language Processing MIT Press, Cambridge, MA.
43. Rogers, JE and Rector, AL 1997 Terminological Systems: Bridging the Generation Gap. Annual Fall Symposium of American Medical Informatics Association. Nashville TN Hanley & Belfus Inc. Philadelphia PA: 610-614.
44. Rumelhart, David E., McClelland, James L., and the PDP Research Group. 1986 Parallel Distributed Processing: Explorations in the Microstructure of Cognition MIT Press, Cambridge MA
45. Alan Saltzman 1985. Adverse Reaction Terminology Standardization: A Report on Schering- Plough's Use of the WHO Dictionary and the Formation of the WHO Adverse Reaction Terminology Users Group (WUG) Consortium. *Drug Information Journal* 19:35–41
46. D. Sankoff and JB Kruskal Time Warps, String Edits and Macromolecules. Reading, Mass.: Addison-Wesley, 1983
47. Fred Schneiweiss Adverse Reaction Thesauri Used in the Pharmaceutical Industry 1987. *Drug Information Journal* **21** 299–302
48. Hinrich Schütze, David A. Hull, and Jan O. Pedersen. 1995. A comparison of classifiers and document representations for the routing problem. In *Proceedings of SIGIR95*, pages 229–237.

49. Ken Thompson 1968. Regular expression search algorithm. *Communications of the ACM* **11**: 419–422
50. Alan Turing 1950 Computing machinery and intelligence. MIND (the Journal of the Mind Association), vol. 59, no. 236, pp. 433-60, 1950
51. Wayne M. Turner, Julie B. Milstien, Gerald A. Faich and George D. Armstrong 1986. The Processing of Adverse Reaction Reports at FDA. *Drug Information Journal* 20: 147-150.
52. National Library of Medicine ULS Knowledge Sources, Unified Medical Language System 13th Edition Janary Release 202AA Documentation
53. Lynn E Webb 2000 Advantages and Disadvantages of Translation Memory: A Cost/Benefit Analysis. MA Thesis, Monterey Institute of International Studies.
54. http://www.who.int/library/historical/access/disease/index.en.shtml WHO Library Historical Collection Jacques Bertillion Nomenclatures des maladies (statistique de morbidité - statistique des causes de décès) arrêtées par la Commission internationale chargée de reviser les nomenclatures nosologiques (18-21 août 1900) pour etre en usage à partir du 1er janvier 1901
55. http://www.who.int/library/historical/access/disease/index.en.shtml WHO Library Historical Collection International Commission for the Decennial Revision of Nosological Nomenclature] Commission internationale pour la révision décennale des nomenclatures internationales des maladies, causes de décès, causes d' incapacité de travail, devant servir à l' établissement des statistiques nosologiques (classification Bertillon) quatrième session, 16-19 octobre 1929. 4e rév. Paris : Impr. nationale, 1930.
56. International Commission for the Decennial Revision of Nosological Nomenclature Nomenclatures 1940. Internationales des causes de décès 1938 (classification Bertillon): cinquième révision décennale effectuée par la Conférence internationale de Paris du 3 au 7 octobre 1938. La Haye, Institut international de statistique. available at http://www.who.int/library/historical/access/disease/index.en.shtml, accessed March 12, 2002
57. World Health Organization 1965. *Manual for international Classification of diseases, 7th revision.* Geneva, Switzerland.
58. World Health Organization 1969. *Manual for international Classification of diseases, 8th revision.* Geneva, Switzerland.
59. World Health Organization 1992. *Manual for international Classification of diseases and health related Problems, 10th revision.* Geneva, Switzerland.
60. Wojcik, Richard, and Hoard, James 1995 Controlled Languages in Industry. Survey of the State of the Art in Human Language Technology, ed. Giovanni Battista Varile and Antonio Zampolli, 274–276.
61. WONCA International Classification Committee 1998 *International Classification of Primary Care, Second Edition* Oxford University Press, Oxford, England.